CASL

The tongue and lips in Lombard speech:

A pilot study of vowel-space expansion

James M Scobbie,

Joan Ma and Jo White

Working Paper WP-21

September 2012

Queen Margaret University

CLINICAL AUDIOLOGY, SPEECH AND LANGUAGE RESEARCH CENTRE

Update and Sourcing Information September 2012

This paper is available online in pdf format 2012 onwards at

http://www.qmu.ac.uk/casl
http://eresearch.qmu.ac.uk/3601/

Author Contact details:

- 2012 onwards at jscobbie@qmu.ac.uk

Subsequent publication & presentation details:

This is a written-up version of a poster presented at the Listening Talker workshop in Edinburgh, May 2012. This working paper should be cited, not the poster.

The tongue and lips in Lombard speech:

A pilot study of vowel-space expansion

James M. Scobbie, Joan Ma and Jo White

CASL Research Centre, Queen Margaret University

Abstract

We investigate some ways in which speech production alters to make speech sounds more intelligible to a listener. This single speaker pilot study uses ultrasound tongue imaging and videos of lips to investigate the underlying articulatory processes used to distinguish six different monophthongal vowels in Scottish English in a consistent b__p frame. Public-loop Lombard speech was elicited in an interactive task with a neutral condition and a condition where the listener's hearing was masked by speech babble in a natural manner with feedback of perceptual errors being given clearly and in real time to the speaker. As a baseline, the acoustic formant space was measured, which showed Lombard effects of F1 lowering for all vowels except /i/ and an increase in intensity. In articulation, we found that in the low and back vowel targets, the hyper-articulated version has extra lowering. However, for high front vowels /i/ and /e/, the hyper-articulated version has slight blade lowering and dorsal retraction in association with raising into the palate. The vowel /ʉ/ has very little change, but seems to fit with the high front set. Lip protrusion and spreading are enhanced, appropriately. Despite the frame being identical in each word, qualitatively the speaker enhanced the /b/ but not the /p/, supporting models in which a CV unit is planned holistically in speech production.

Keywords: Lombard speech, hearing impairment, vowel space, human speech production, ultrasound tongue imaging, UTI.

1 Introduction

Various contexts, tasks and motivating factors cause speakers to alter aspects of their speech production in order to enhance the distinctiveness of a word in comparison to the cohort of similar words, a single competitor, or something in between. Such lexical enhancement can be viewed as maximising the phonological contrast between the specific segments that serve to distinguish the word from neighbours in the cohort (Lane and Tranel, 1971), perhaps weighted towards those contrasts that bear the greatest relevant functional load. Enhancement in that view should be primarily local and unequal rather than lexical (or continuous) and holistic: it is achieved by enhancing the clarity of some or all of the constituent segments of a word in the context of some sort of confusion matrix. Such segmental enhancement might target all the segments in a word; perhaps all of them in the default case or perhaps

contingent on each segment functioning to distinguish the word from its neighbourhood. Even with the smallest possible neighbourhood or where a unique segmental target is enhanced, the phonetic ramifications might spread out to affect the whole word. Alternatively, this sort of phonologically functional enhancement can be primarily holistic, achieved through clarifying the lexical identity of a word via global changes; ones applying to whole utterances, such as reducing speech rate or increasing overall intensity.

In either case, multiple factors are in play. The mutual distinctiveness of words which comprise a confusable minimal pair, like "fin" /fɪn/ and "thin" /θɪn/ can be enhanced both by making the contrasting phonemes more dissimilar to each other, and by making each of the initial fricatives closer to some global canonical linguistic target – so that in this example the /f/ would become more [f]-like and the /θ/ more [θ]-like. These are not identical strategies. Moreover, the latter conception assumes there is a uniform and transparent relationship between a phoneme and its phonetic target as symbolised/phonologised. The former is less unrealistic, because it seems clear that any phonemic contrast can be increased in multidimensional phonetic space, but it is not deterministic: there are various routes to such enhancement. It is also important to bear in mind that the mutual distinctiveness of whole words is not necessarily enhanced in perception by maximising the closeness of each segment to some purported target, nor by maximising (for each segment in the word) the contrast between that target segment and one or all of its competitors. Nor is it at all likely that the same enhancements would be equally effective for all listeners. Indeed, some may be counter-productive.

In sum, the perceptual enhancement of a word will be embodied phonetically in speech production, and can be expected to vary locally with respect to the inherent characteristics of specific speech sounds, with respect to the potential confusability of the word in a cohort of competitors, and with respect to the various communicative functions concerned.

One of the specific functional reasons to produce clear speech is to overcome a problematically attenuated signal-to-noise ratio, whether located in the human perceptual system or in the environment. The Lombard effect originally concerned the production of increasingly loud speech in an increasingly noisy environment. It is canonically assumed (Lombard, 1911) to be automatic and reflex-like and expressed through general factors such as increased intensity and duration of speech sounds, but may also involve segment-specific or contrastive factors such as the use of an expanded vowel formant space. It has long been known that speech elicited in such noisy environments is indeed more intelligible (e.g. Dreher and O'Neill, 1957).

However, there are a number of overlapping phenomena in this area, including variable importance of factors that motivate clarity, and it seems appropriate to assume that in many real world situations both automatic and phonological neighbourhood effects will be relevant (Lindblom, 1990; Junqua, 1996; Garnier et al., 2006ab; Nicolaidis, 2012). Additionally, it is important not to forget that what is "clear" for one speaker or hearer may not be for another, linguistic systems being so heterogeneous within any well-structured speech community let alone in more

unpredictable interactions. Moreover, social rather than lexical meaning may be the aspect requiring primary clarification, and so systematic sociolinguistic variation within a speech community or style factors and important factors (Uchanski, 2005; Smiljanić and Bradlow, 2008). In many circumstances, for many linguistic varieties, "clear" speech styles involve using socially-distributed variables, and awareness of the role of standard or prestige targets (see Wassinck, Wright and Franklin, 2007).

It seems particularly important to take this complex interaction into account examining the more linguistically-structured effects of noise on speech, because different aspects of clear enhanced speech may be attributable to different parts of the multiple interacting systems involved in structured linguistic variation. A vernacular production of "thin" as [fɪn] which could be mis-heard as "thin" could be changed to a more standard [θɪn] in some sociolinguistic contexts in order for the speaker's lexical choice to be more clear, whereas in other contexts the same speaker could be clearer in their lexical target ("thin") by increasing the intensity of the [f] in [fɪn]. In yet other contexts, it may be that social meaning attaching to the use of [f] is prioritised over lexical clarity. Thus sociolinguistic variation can make lexical targets clearer not just through subtle changes in phonetic target, but by altering the linguistic target categorically towards a different system that is, perhaps, more in line with listener expectations or experience. But the speaker does not need to become more standard. Whether it is reasonable to characterise these options for clarification as a single, abstract or multidimensional system is less clear, though Wassink, Wright and Franklin (2007), for example, combine various aspects together under the concept of audience design.

Returning to a more basic phonetic concern, perceptibly clearer speech sounds have to arise from changes in speech production. They might include global speech characteristics (like effort or speech rate), prosodically focussed ones, and/or (in a context-sensitive manner) changes to the specific segmental characteristics of vowels and consonants. The phonetic characteristics of such segmental "hyperspeech" that we look to has been exemplified in a number of acoustic studies of vowel space expansion (e.g. Johnson, Flemming and Wright, 1993), and in an articulatory study of consonants (Nicolaidis, 2012).

From the point of view of articulation, Nicolaidis's innovative electropalatographic study of consonant production found a set of trends towards the use of greater articulatory effort (greater linguo-palatal contact, less coarticulation and shorter duration) in Lombard speech, though statistically significant differences were often unattested among these general trends. The duration of contact was significantly shorter in general, but for individual consonants showing a reduction of between 3% to 14% in the Lombard condition, post-hoc tests were mostly not significant, as, for example, in the cases of /t/ and /k/, which averaged 108ms and 99ms respectively in the neutral condition, while in the Lombard condition they were 105ms and 93ms.

Greater effort and the associated greater acoustic intelligibility have been associated with larger facial movements (e.g. Sumby and Pollack, 1954; Vatiokis-Bateson et al., 2007). These increased movements of the face are presumably associated with the underlying speech movements in the jaw and tongue (as well as

the lips) which are in part causally responsible for some of the spectral expansion of the vowel space as well as increased intensity and duration. Vatiokis-Bateson et al. also show how these visible changes are part of increased clarity when the visual modality of audio-visual speech perception is taken into account.

Overall, we therefore tend to the view that it is unlikely there is a neutral, physical reflex explanation for the full set of changes that can be attributed to increasing the clarity of speech, even in a simple case. This is partly due to the varied phonetic characteristics of different speech sounds and the functional load that individual phonological contrasts bear. But it is also due to the social embedding of all linguistic systems, in which some aspects of the system are regarded as being more distinctive, canonical, or standard than others. Socially, some accents are regarded by some speakers as having greater inherent intelligibility than others, and some are more familiar to some of the participants in a discourse.

Here, however, we are more concerned with speech production itself, because it is not possible to alter the production of just one aspect of speech without affecting others. It is hard to model simple, universal enhancements in articulation: making speech clearer is quite unlike turning up the volume on a recording. How even global increases to articulatory effort are implemented in speech is an interesting, complex, and little-addressed question (Nicolaidis, 2012). It is not possible for example to alter one phoneme without altering (its relationship to) others, and not possible to alter the production of one phonetic correlate of a phoneme without ramifications of this on the balance of cues.

Greater intensity, for example, is partly enabled by a greater opening of the mouth, in which case vowel quality is affected too: *all* vowels should have a lower jaw and tongue position, with F1 (the first formant) for all vowels ending up with a higher frequency. But this would translate, rather than expand, the vowel space, whereas an expansion would suggest differential degrees of jaw opening and tongue lowering appropriate to phonological vowel height. If speech rate is slowed and voiced portions given greater proportional duration, then co-articulatory and prosodic changes will occur, not just a simple stretching, making the consonants that flank the vowel less canonical in their transitions. Nicolaidis (2012) for example found less vowel coarticulation onto consonants.

Added to these multi-dimensional interactions at a phonetic level, the complexity of the psycholinguistics of speech perception (even drawn narrowly in relation to lexical intelligibility) should not be underestimated. In the real world, the functional considerations relating to our aim to successfully transmit meaning ranges over various linguistic levels, as indicated above. It may never really be the case that just lexical/syntactic information is affected, to the exclusion of affective, idiolectal, sociolinguistic and discourse-oriented ones. Given physiological differences between speakers, we must also assume that all things being equal, systematically identical speakers in identical contexts will still aim for efficient, clear speech production in slightly different ways.

It therefore seems clear that very many factors other than just intensity are, or can be, manipulated by speakers as part of a wide set of different types of “speaking

clearly". Thus we need to recognise that this is why a number of relevant partial models exist in relation to clear speech, evidenced and explored in a range of different research traditions. For example, one perspective familiar to most linguistic phoneticians is the rather biologically-inspired work by Lindblom that adopts a holistic hyperspeech/hypospeech dimension (Lindblom, 1990). Proponents of any such model presume that it can be expanded to encompass other influences on, and modes of, speech production relevant to considerations of intelligibility and clarity, broadly construed.

Our interest here is on speech production – to what extent is it useful or revealing to explore clear speech from an articulatory perspective?

We ask this because most work on clear speech has used perceptual methods to probe intelligibility directly, augmented with acoustic analysis methods to probe speech production. Using such research, inferences can indeed be drawn about speech articulation itself (e.g Picheny, Durlach and Braida, 1986). However, given the complex links between acoustics and articulation, and given that very little work has directly explored intra-oral articulation of clear speech in phonetic depth (with the notable exceptions of Matthies, et al., 2001; Nicolaidis, 2012), little is known about how speakers subconsciously alter their lingual articulations to achieve their ends, e.g. in Lombard speech, though there are some clear general observations. Based on observations of jaw lowering as well as acoustic changes in F1, we might expect a vowel's tongue surface to be lower in clear speech than in the neutral condition, and for low vowels to be affected more than high ones, in order to expand the acoustic vowel space. From a dynamic perspective, slow-clear speech might show faster transitions from target to target with longer stable targets which are not subject to undershoot and whose targets coarticulate less with the context.

While a number of quantitative techniques are available for articulatory phonetic research, such as MRI, EMA, motion capture, EPG, few have been put to use in this area (though there is body of work on localised linguistic enhancement and the prominence of specific words). Two accessible technologies that could reveal a great deal about more generalised enhancement and clarity in speech production are, first, Ultrasound Tongue Imaging, which is well suited for the measurement of the shape and location of the tongue surface in the mid-sagittal plane, particularly useful for characterising vowels. Second, video camera recordings can be used to explore differences in the articulation of the lips. In both cases, the location of the measuring equipment relative to the speaker's head needs to be kept constant, or to be compensated for, which is awkward both for data collected via moving camera or for stabilised cameras where the speaker themself is free to move their head naturally.

The articulatory measurement of clear speech could be applied in a number of areas, such as silent speech interfaces, communication with listeners with a hearing impairment, or forensics. Moreover the interpretation of variation (in its extremes, modes and dimensions) is valuable in phonology and sociolinguistics, providing insight into distinctive or salient features of contrast and structure. Compared to other articulatory techniques, ultrasound and video have potential as portable, cheap and accessible technologies. Since cross-linguistic and cross-dialectal differences in

system are so important, it is like that larger scale studies will use this kind of methodology, if it shows promise.

Our main research questions in this pilot study, which will focus on vowels in Lombard speech in a single Scottish speaker of English, are therefore:

- Methodologically, can ultrasound and video capture key aspects of clear speech articulation qualitatively and quantitatively?
- Are the articulatory aspects of clear speech interestingly different from traditional acoustic measures?
- What do the results tell us about Scottish English?

2 Method

2.1 Participants

The single speaker was an adult female, aged between 20-30, with a Scottish accent of Standard English. The single listener was an adult male, aged between 30-40, with a northern English accent. Both were postgraduate students in Speech and Hearing Sciences at QMU.

2.2 Materials

Six representative vowels of the nine monophthongs of Scottish English were elicited, namely /i e a ɔ o ʉ/ (which could be said to be the six bimoraic monophthongs). The vowels were placed in a labial /b__p/ context (creating a mix of real and pseudo-words like /bap/ and /bep/) to avoid competing lingual specifications from the consonants onto the vowel, though closing a syllable with a stop means the vowels were phonetically rather short. Only open mono-syllables which are onsetless (or with an /h/ onset perhaps, like /hu/) could be less specified for consonantal lingual targets – we would expect an open syllable allophone of the vowel to be much longer in duration, and, like so many other factors not examined in this experiment, we'd expect this to alter the details of the Lombard effect. The materials were represented orthographically as: *beep, bape, bap, bop, bope, boop*.

2.3 Protocol

Ten tokens of each word were presented on screen using AAA software version 2.14 (Articulate Instruments, 2012), randomized in a single block (n=60) that did not disallow sequences of identical words. The block was elicited first in a neutral condition, then with a different single block randomization, in a Lombard condition (total n=120).

Speaker, listener, experimenter and some observers were present in a sound-treated recording room. The materials were presented orthographically one word at a time, on screen, to the speaker. Each condition lasted about 12 minutes. In both conditions the interactive listener, who was seated about 1.5m from the speaker, could see the speaker's face and hear them speak, but not see the prompts. The listener knew their task was to repeat the word correctly back to the speaker, who would confirm its correctness, and that all words were one of six similar monosyllabic words. If the listener's response was zero or incorrect, the speaker

indicated a failed response, and then repeated the word to give the listener a second attempt. We did not analyse the distribution or nature of incorrect tokens, just the ones that were repeated correctly, but there were approximately one dozen mis-hearings, almost all in the second condition.

In the first, neutral condition, the listener's hearing was unimpeded and he wore no headphones. In the Lombard condition, the listener wore headphones and his hearing was masked by speech babble at 60dB, making mishearing much more likely. However the point of the experiment was for the speaker to avoid this situation, the task eliciting instead clear speech from the speaker as they subconsciously adjusted their speech to the situation.

Both speaker and listener knew that there would be two conditions before the experiment started, and in the appropriate order, and both knew the rough purpose of the ultrasound and acoustic measures.

2.4 Data capture

AAA software (version 2.14) was used for data capture. Ultrasound images were obtained from the QMU Ultrasonix system, the probe mounted on the Articulate Instruments Ltd stabilising headset (Articulate Instruments, 2008). Each scan from the probe resulted in a unique data frame, with no compression of data at storage time. The sample rate was 120.7fps (frames per second). Images of the lips were taken from a headset-mounted video camera, de-interlaced to 59.9fps on the assumption the camera was operating at the NTSC standard rate. The field of view was 134° and the probe frequency was 5MHz. There were 63 scanlines of 412 bits, with a theoretical separation of 2.14° (assuming each to be a linear beam).

UTI analysis typically takes as input a real-space image which represents a midsagittal fan-shaped "slice" of the vocal tract within a rectangular frame. The visible image is usually subject to image processing within the ultrasound scanner to create video-streamed output, and may then be digitised and stored (e.g. in avi files) with compression. Even where the raw data that created this fan is available for storage (as here), without such additional AD effects on spatio-temporal accuracy (Wrench and Scobbie, 2006), the resolution of the image can be hard to calculate, and it varies.

Here, resolution is a little easier to estimate because in the QMU high speed system each probe scan results in a single image frame, whatever the frame rate, and there is no buffering of data to create a composite image, and the fast scan rate minimises error in . The circumferential resolution of UTI varies depending on the distance of the tongue surface from the probe (because data between radial scanlines is interpolated), and the narrowness and of the original scanline, which is not in fact linear but is a spreading wedge of energy. The raw axial radial resolution varies depending on the depth, number of bits used for a scanline, and number of pixels available to code the bits.

For this experiment, standard settings for the QMU laboratory were used: the theoretical angular (circumferential) resolution at a 6cm distance from the probe

surface is approximately 2.6mm. The axial (radial) resolution along a scanline is about 0.5mm.

In addition to recording active articulators, a palate trace was made using gentle dry swallowing, and an occlusal plane recorded with a plastic biteplate (Scobbie et al, 2011). With appropriate co-registration of video and ultrasound, the external and internal images can be spatially aligned as well as being temporally synchronised, but this was not done here.

2.5 Acoustic analysis

F1 & F2 averages were measured in each vowel, extracted in PRAAT from a variable section of the vowel (avoiding formant changes due to initial and final formant transitions from the labials) selected to be as long as possible, but avoiding any clipped central portions of the waveform. (The speaker's Lombard speech was so high intensity that clipping of acoustic data occurred in most tokens early in the vowel.) Values were measured in Hz, and converted to Bark (Traunmüller 1990) using the formula

$$\text{Bark}=((26.81\times\text{Hz})\div(1960+\text{Hz}))+0.53$$

in order that a unit difference in Bark is comparable at every frequency, and in order that we can present fixed-aspect ratio diagrams of vowel space that show equal-auditory-value movements in both directions. The vowel space area can be therefore measured, in squared Bark. Triangles are used for this approximation though the images below have smoothed area perimeters for aesthetic reasons.

The duration of each acoustic "word" was also measured (Figure 1). We also report the ratio of the vowel to word as a percentage. The word is defined as "V" starting at the /b/ burst including voiceless portions up to the voice onset (i.e. *including* any non-zero voice onset time), the voiced portion of the vowel, plus "T", any short final transition (weak energy / devoiced / ambiguous / echo) which could be indicative of closure, plus "C" the clearly silent closure phase of the final /p/ up to but not including its burst energy. The final burst was always detectable, though sometimes very quiet, but the transition was rather variable, so the acoustic measurement of the final consonant duration as either C or T+C might be rather problematic. VOT was generally short and often indistinguishable from zero (as in the case illustrated). The word duration here is effectively the acoustic rime duration.

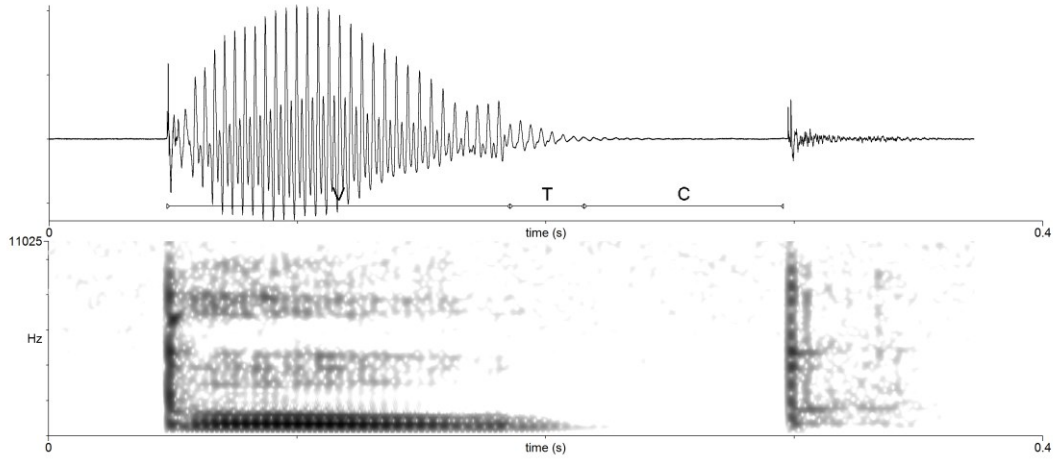


Figure 1 Acoustic Annotations of V (vowel), T (transition) and C (consonant closure) in a typical token (neutral_bope_2)

We were not very confident that annotations to separate vowel, transition and the final silent closure would show high levels of inter-rater or intra-rater reliability, or that such acoustic measures would be highly replicable across speakers. We will also report some articulatory durations (see below), but due to synchronisation problems, no attempt will be made to relate the timing of articulatory events to acoustic ones.

2.6 Articulatory analysis

AAA was used for analysis as well as data capture. Semi-automatic tracing of mid-sagittal tongue surface was performed with a single frame being extracted from a stable midpoint in the vowel. The tracing was made on AAA's 42 radius measurement fan, so that each tongue curve has 42 control points representing the distance from the mid-sagittal tongue surface from the centre of the probe along that angle of incidence reflecting the directions of the scanline/echo-pulse beams from the probe.

For each condition, the ten tokens were averaged using the AAA averaging function in the AAA workspace and plotted as a mean point on each fan line, and flanked by +/- 1 s.d. i.e. with a larger or smaller radius. To produce an image of the shape of the tongue surface in the mid-sagittal plane, these points can be joined (by a spline), in a diagram with a fixed-aspect ratio of 1x1. If an articulatory image has a different aspect ratio, then the tongue shape is needlessly distorted and is harder to interpret qualitatively or to compare to figures from other studies. By default, a true ratio should always be used.

Significant differences were tested along each fan line with the built-in t-test facility of AAA. A significant value for a single t-test (i.e. along one fan-line) is unlikely to be of linguistic interest, however low the p value: tongues are not spiky and one single significant value of multiple sigmas should not be given undue importance. Instead, a relatively long mid-sagittal region of tongue surface with repeated significance at p=0.05 is more likely to indicate a meaningful difference in constriction between conditions, even though the closeness of the fan-lines certainly

means that the t-tests are not independent, so ideally should have some kind of statistical correction to avoid Type-1 errors (false positives).

To help avoid a Type-1 error, we will only interpret as significant a contiguous series of 5 or more significant ($p<0.05$) radial differences on adjacent radial fan-lines, rather than correcting individual p values. Requiring a threshold of multiple adjacent significant values is the method of correction.

The actual distance between the two tongue curves on each of these significant radial fan-lines is available in AAA (along with non-significant differences): these contributed to a mean distance (the sum of radial differences divided by the number of radii concerned) between the two tongue curve means. And, taking into account the distance of the tongue from the probe along each radial fan-line, we estimated the area between the tongue curves for each single sector defined by a radial fan-line. Similarly, it is possible to report the length of the tongue surface to which this zone of significance pertains. In addition, given that the two tongue curves can cross over in a sort of X pattern, necessitating a zone with no significant difference in distance from the probe along the fan-line, it makes sense to include these small or zero differences in radial difference in the calculation of mean difference between the surface curves (even though they cannot establish significance in the first place). We will thus report the average distance between the curves, and the total area of difference, including non-significant fan-lines that are contiguous with significant fan-lines in cases like those just mentioned, specifically cross-over.

The reason we adopt the heuristic of 5 continguous fan-lines is based on a simple observation that while adjacent and close points on the tongue cannot act independently, points far apart on the tongue certainly can. A quick correlational analysis suggests the high and significant correlation between adjacent parts of the tongue tends to fall off rapidly at a separation of about 5 fan-lines. A firmer statistical model, looking at the front, middle and root of tongue separately would provide a clearer heuristic, but this work remains to be done.

So, unlike SS-ANOVA (Davidson, 2006), AAA's fan averaging estimates significance bottom up, one thin slice of the measurement fan at a time. This means that our positing of a significant difference between two tongue contours/curves

- is clearly located at particular points on the tongue surface, and
- must be of a minimal length of tongue surface

The interpretation of SS-ANOVA's non-overlapping confidence intervals as indicating different tongue surface shapes or locations could, likewise, require a minimum length of tongue surface. And, this use of bottom-up t-tests could make use of top-down information to focus on, for example, tongue contour differences in particular parts of the oral cavity.

The area between two tongue surfaces was calculated as the sum of annular sectors bounded radially intermediately between fan-lines and in its outer and inner circumferences by the fan-line distance of the tongue surface from the probe.

Lip compression, aperture, protrusion and spreading were estimated rather more speculatively direct from the video images, primarily using the length of a straight line, namely the shortest common tangent across both lips (Figure 2, Figure 3). This distance is rather less than a fleshpoint-to-fleshpoint difference between upper lip and lower lip (e.g. between EMA coils UL-LL in 2D or between two motion-capture fleshpoints), depending on the size and shape of the lips, but is comparable.

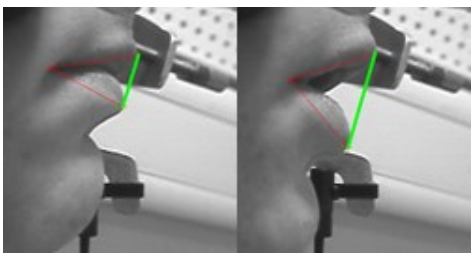
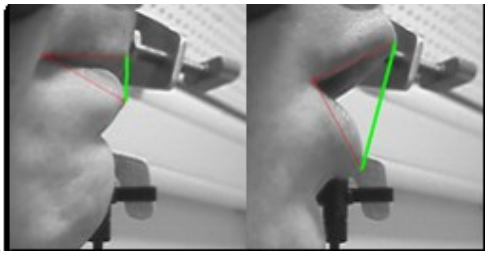
a.  b. 

Figure 2 Lip closure/opening measurement vector (green line) and lip spreading vector (lower red line) showing /b/ and /ɔ/ (bop) in (a) the neutral condition and (b) the Lombard condition.

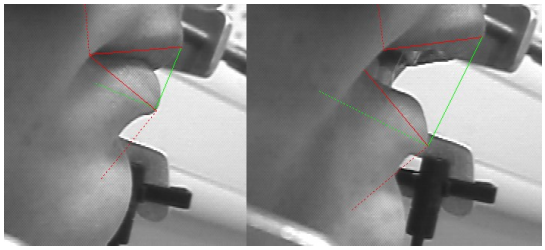
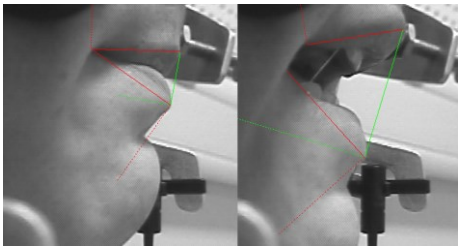
a.  b. 

Figure 3 Lip closure/opening measurement vector (green line) and lip spreading vector (lower red line) showing /b/ and /a/ (bap) in (a) the neutral condition and (b) the Lombard condition.

A variety of ways were attempted to locate the lateral margins of the lips to estimate lip spreading, but these were regarded as unsuccessful, due to a nasolabial fold which obscured the laterial of the mouth when strongly retracted or even in a neutral closure, in part due to the headset cheekpad compressing the cheeks. Any kind of fleshpoint marker (e.g. a pen-mark) or small attachment to the skin would have given much more replicable findings but still would not have captured actual closure without interference. Blue lipstick and colour-based analysis building on Liptrack (Lallouache, 1991, cited by e.g. Noiray et al., 2011) would, we think, be preferable.

We do report here the length of the lower lip linear measure for the vowels, but interpret it only as spreading for the unrounded vowels (Figure 3) and do not recommend its use. If the edge-of-mouth had been visible, its location in the sagittal plane could have been plotted. We can and will however plot the location of the tangent termini as proxies for UL and LL locations, but only for qualitative discussion. Measures of protrusion, closure duration, compression and opening based on linear analysis and manual direct video tracing do not appear to be so problematic, though were undoubtedly inefficient to undertake, and appear suitable for quantitative analysis.

When the lips were closed but apparently barely touching (Figure 1a, left image), the pixel-based distance (10 arbitrary units) was re-based for clarity to be 0 units. Positive measures therefore represent a distance compatible with the lips being parted

with an open airway (Figure 1a and 1b, right images), while a negative value is much more likely to indicate closure of the airway, with more extreme lip compression leading to lower negative values (Figure 1b, left image). In the examples here, in Figure 1 (bop) the normalised lengths of the green tangent lines are -2, 4 and -4, 8 units for neutral and Lombard conditions, and in Figure 2 (bap) -1, 6 and -2, 11 units.

A final limitation on these methods is the camera angle. The camera should be vertically closer to the plane where the lips meet for protrusion and compression measures, and should be imaging the coronal place rather than the sagittal plane for lip spreading.

2.7 Statistical analysis

Given that this is a pilot study with a single speaker, there is only limited statistical analysis. Most results are stated in the manner of qualitative statements of descriptive statistics. It is more important for exploring these results that more speakers and more segmental effects are explored, rather than enhancing the statistical analysis within this pilot study. Some exploratory use of t-test and ANOVA was undertaken, to help decide what results should be given particular attention, so reference to "significant" results should be understood in this context. For the articulatory data, even for a single speaker it is, however, appropriate to report significant differences in tongue surface contours, using methods described above.

3 Acoustic Results

3.1 Impressionistic comments

In general qualitative terms, the Lombard condition was a lot louder than the neutral condition, but not necessarily much clearer. Given the clipping of the recordings, these comments were based on live impressions, and re-listening is pointless, as is any use of the recordings for perceptual analysis. There was impressionistic lengthening of word duration. The initial stop /b/ sounded more energetic, and the final /p/ seemed relatively quiet in the Lombard condition.

3.2 Duration

In terms of vowel duration (Figure 4), the Lombard condition is consistently longer for all vowels, by 38% on average (s.d. 10%), with inconsistent indications that phonological vowel height operates as a factor, so that low vowels have a greater duration. The smallest increases of 26% and 28% are for /ʉ/ and /e/ respectively, the largest (47% and 49%) for /i/ and /a/, with intermediate values of 43% and 34% for /ɔ/ and /o/. The low vowel /a/ is the unique longest in the Lombard condition, and /i/ and /ʉ/ are shortest in both conditions, in line with Scottish English norms and the phonotactic context. The relative durational relationships of the vowels may or may not be preserved / enhanced in the Lombard condition.

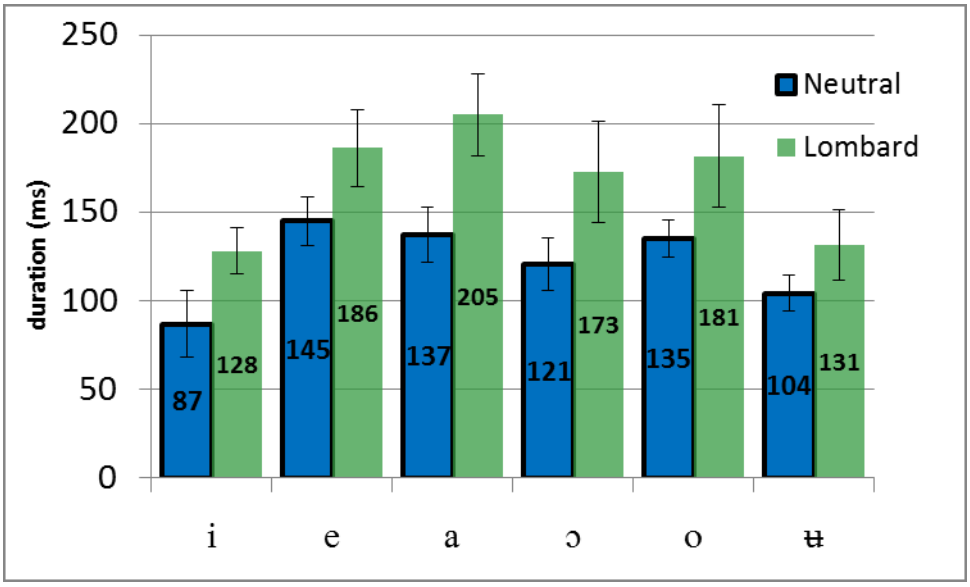


Figure 4 Acoustic vowel duration in both conditions

The vocalic transitional phrase at the end of the vowel is fairly short for all vowels, and quite varied, but increases in all cases in the Lombard condition, from an average 31ms to 55ms (79% on average). Figure 5 shows that the duration of the silent closure phase of the final consonant /p/, on the other hand, is consistent across vowels and shows a stable *decrease* in the Lombard condition of 17ms (range 7-32) from 84ms to 67ms, which on average is -19%.

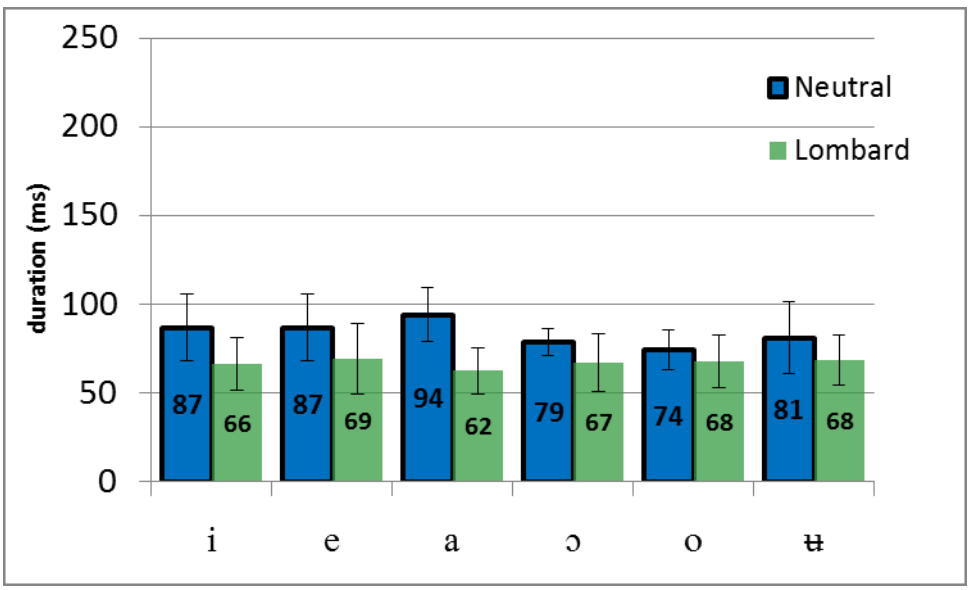


Figure 5 Acoustic duration of final consonant /p/ silence

The ratio of vowel to rime therefore shows an increase in all vowels (Figure 6). The overall effect is significant: the acoustic vowel forms more of the acoustic rime (57% vs. 51%) in the Lombard condition.

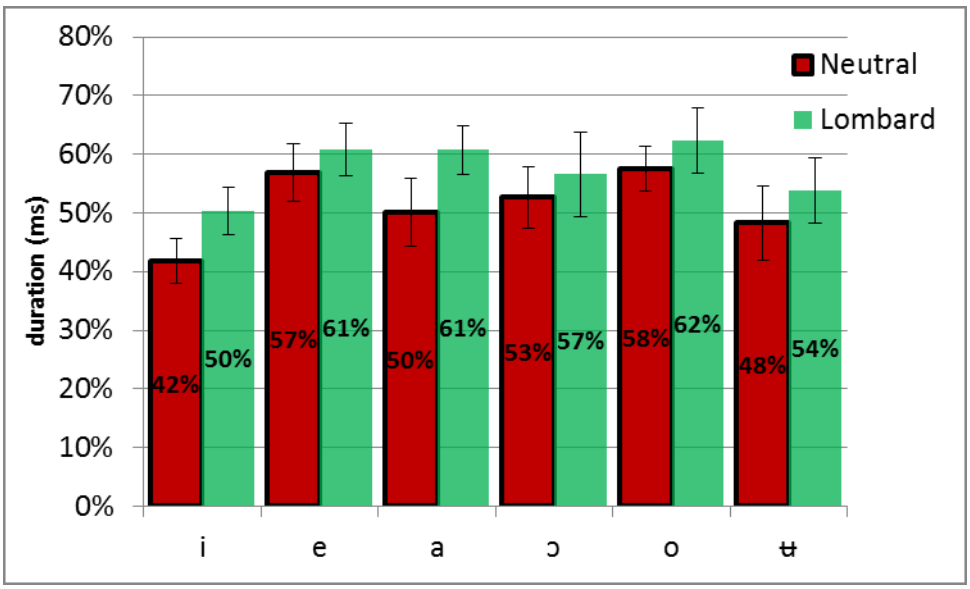


Figure 6 Ratio of vowel to rime duration

While the silence (a clear indication of voiceless closure) is shorter in the Lombard condition (Figure 5), the decreasingly intense transition between the vowel and the silent portion of final /p/ is *longer*, rising from an average 31ms to 55ms, an increase of 71%, but as might be expected, these values are quite variable. Taken together (token-by-token), the low-intensity, non-vocalic portion of the rime comprising both transition and silence is similar in the neutral and Lombard conditions and shows no consistent pattern of increase or decrease in duration across all vowels. Full duration details are in Appendix 1, and see also section on the articulatory duration of /p/ closure.

3.3 Formants

The second formant F2 is mostly stable across conditions, while F1 rises for all vowels bar /i/, suggesting articulatory lowering of the vowel space in the Lombard condition. The other five vowels have with a tendency to raise F1, with on average, a difference of 0.6 Bark (Figure 7). The highest back vowel /o/ and mid-high front /e/ (both monophthongs in SSE) show raised F1 more than /ʉ/ does. In other words, there appears to be a shift down, rather than an increase in size, of the vowel space. Its area in the Lombard condition is indeed very similar to the area in the neutral condition (14.52 $Bark^2$ vs. 14.57 $Bark^2$ respectively).

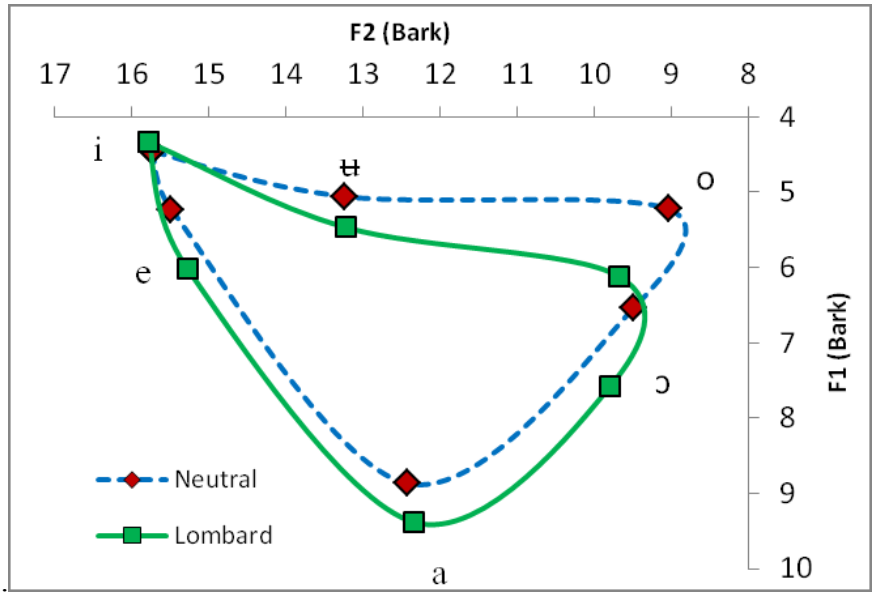


Figure 7 Formant space (fixed-aspect ratio)

We can expect a certain amount of baseline token-to-token variation in the neutral condition. We might expect either a greater amount of variation in the Lombard condition and/or a continuous gradient increase in F1 (not F2, given that it remains relatively constant), given the live feedback to the speaker, and the effect of experience. Qualitatively, F1 like F2 remains relatively constant in both conditions, or shows random changes. We do not see, for example, F1 increasing monotonically or gradually from the start to end of the Lombard condition (Appendix 3).

4 Articulatory Results

4.1 Qualitative tongue surface shape and location

Mid-sagittal lingual tongue curves are all distinguished from each other (Figure 8). Three vowels (/i/, /e/, /ʉ/) have a front/palatal constriction or approximation, while /a/ is low with a retracted tongue root and neutral (fronted and raised) blade. Strong backness/dorsality is shown by /ɔ/ and /o/ with tip lowering and retaction right down into the floor of mouth.

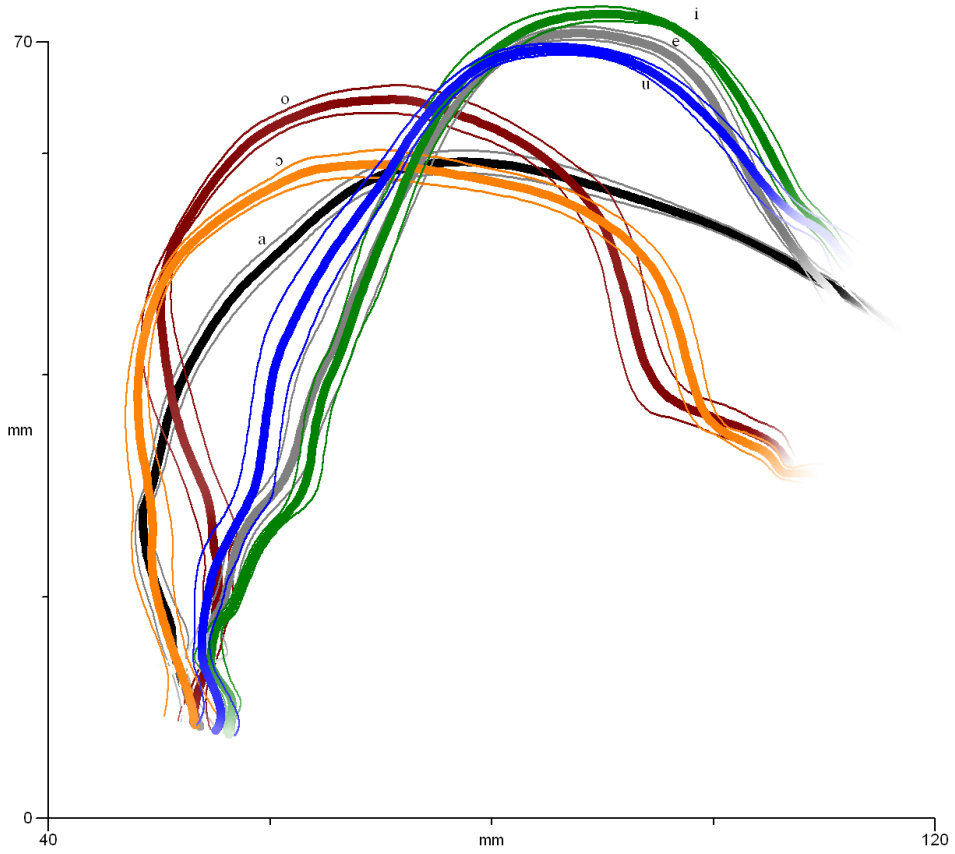


Figure 8 Six vowel means (thick lines) and variation (±1 radial s.d. in colour-matched thin lines) rotated to the speaker's occlusal plane, with tick marks at 20mm intervals. The most anterior part of the traced surface for /o/ and /ɔ/ is floor of mouth, not tongue. Lombard condition.

The fixed aspect ratio of this figure means it can be rotated and measured at any angle without distortion. Each mean is clearly separated from all the others, though /i/ and /e/ overlap and approximate to each other more extensively than any other pair.

Turning now to neutral vs. Lombard comparisons, we see non-overlap in the images in some locations. For each pairwise comparison, where the range of variation indicated by one standard deviation does not overlap there is likely to be a region of significant difference (roughly speaking); actual significance as reported by AAA software for each fan-line was used as the actual criterion for difference, as described above, and the size of difference quantified below. However, first we will discuss the (possible) differences in more holistic terms.

4.1.1 /i/

The front part of the Lombard version of the /i/ of FLEECE is overall slightly retracted, away from the alveolar area to the palatal-dorsal area, with no change in the its root. The blade is a bit lowered, but the rear part of the front is raised, giving the appearance of a backwards movement. There is a cross-over at about (95,70) where

some part of tongue surface occupies the same area in oral cavity space. However, at this cross-over point not only is the orientation of the tongue surface different, there is very likely a different flesh-point at that location.

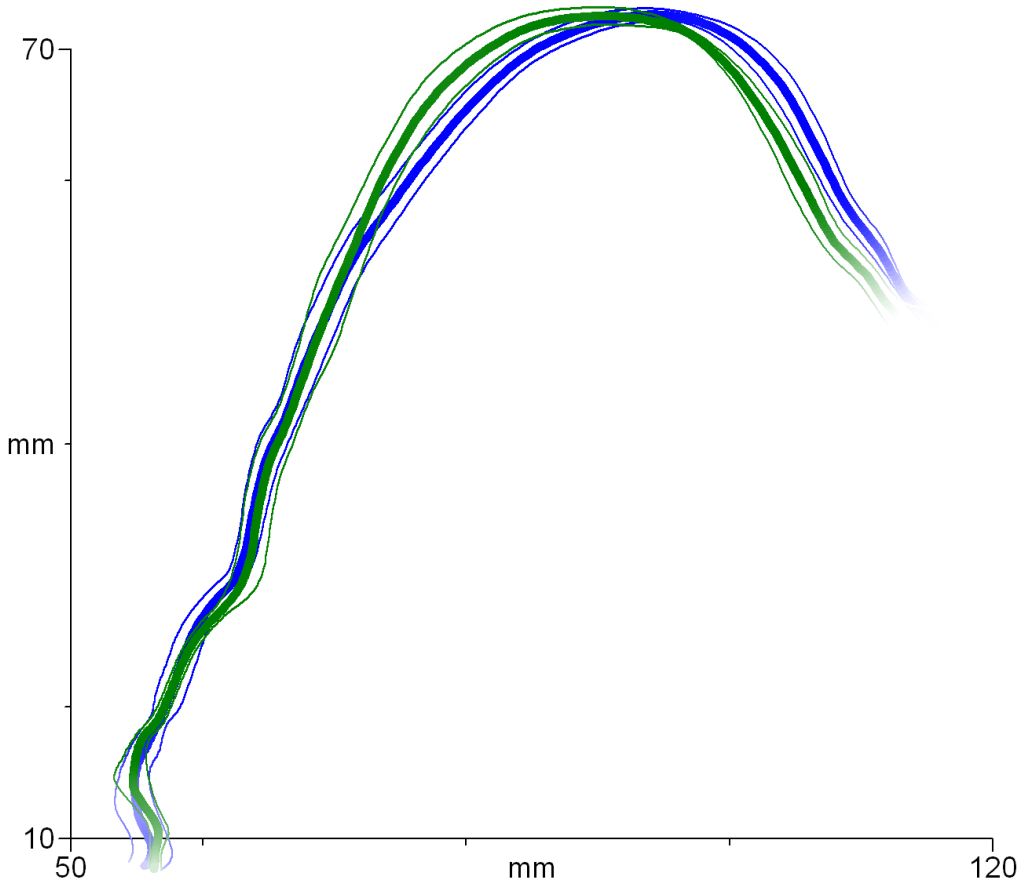


Figure 9 /i/ - with blue neutral and green Lombard (for colour-enabled views)

4.1.2 /e/

The pattern for the /e/ of GOAT is very similar to /i/. Figure 8 above shows that /e/ is a little lower overall than /i/, and the quantificational results below confirm the visual impression that the means are closer together. The conditions appear more similar posterior to the cross-over point than they did in /i/.

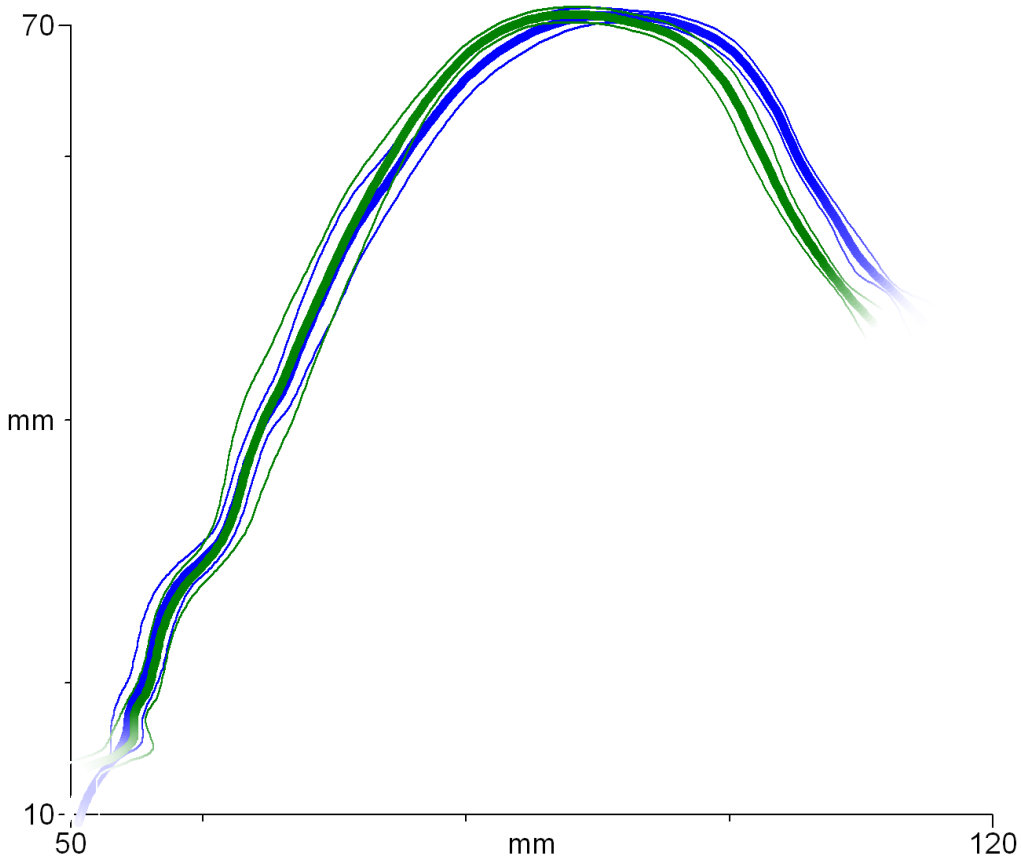


Figure 10 /e/

4.1.3 /ʉ/

Next, we will consider the GOOSE & FOOT vowel (merged in Scottish English). There is almost no indication of a Lombard effect, except perhaps in the same sort of blade lowering seen above. There is, almost invisible, a suggestion of a cross-over point with a very small difference posterior to the cross-over, with a clearer non-difference down the back to the lowest part of the root, where the data is less reliable, and there is no consistency among these three vowels.

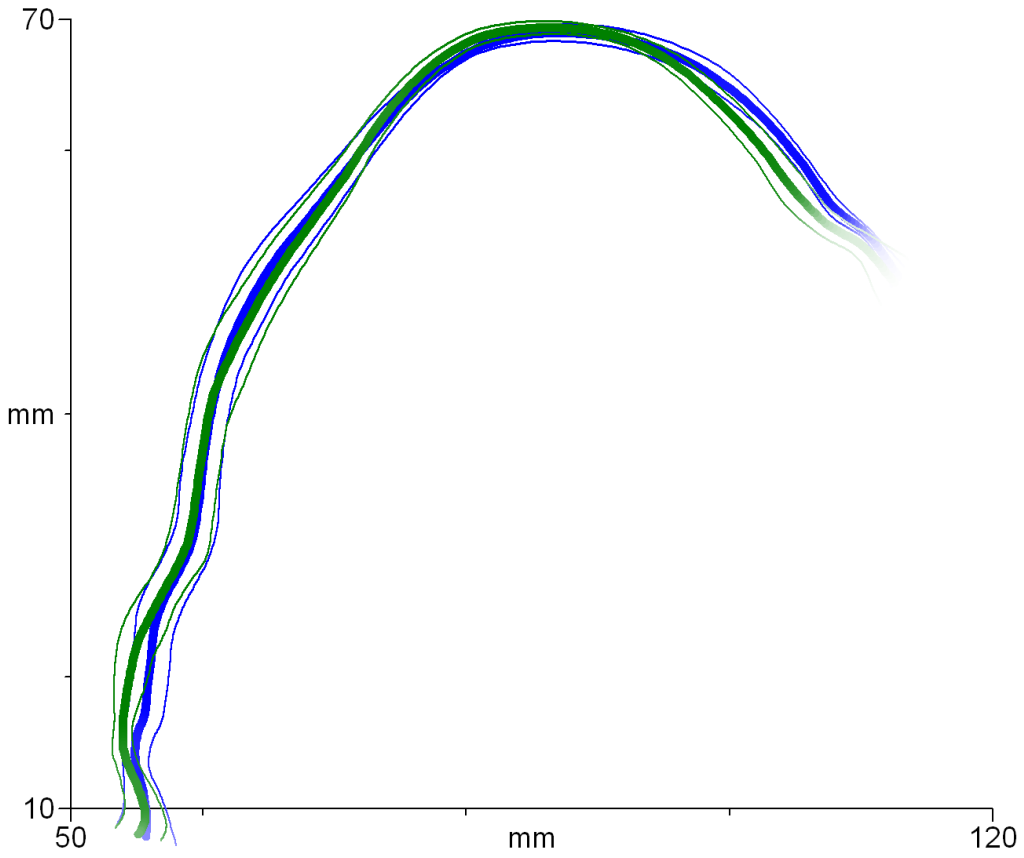


Figure 11 /ʉ/

There is no lowering of the highest part of the constriction (orthogonal to the occlusal plane) in any of these three vowels.

4.1.4 /a/

This vowel from TRAP & PALM shows more of the expected lowering of the "highest" part of the tongue, in the front opposite the hard palate. There is no indication of lowering at the tip, which is probably raised a little off the floor of the mouth in both conditions, not that this can be seen directly in the images. The back and root perhaps show a very small consistent difference in the surface in the Lombard condition, being fronter. (This nowhere reaches significance for a long-enough distance of tongue surface so is not quantified below.)

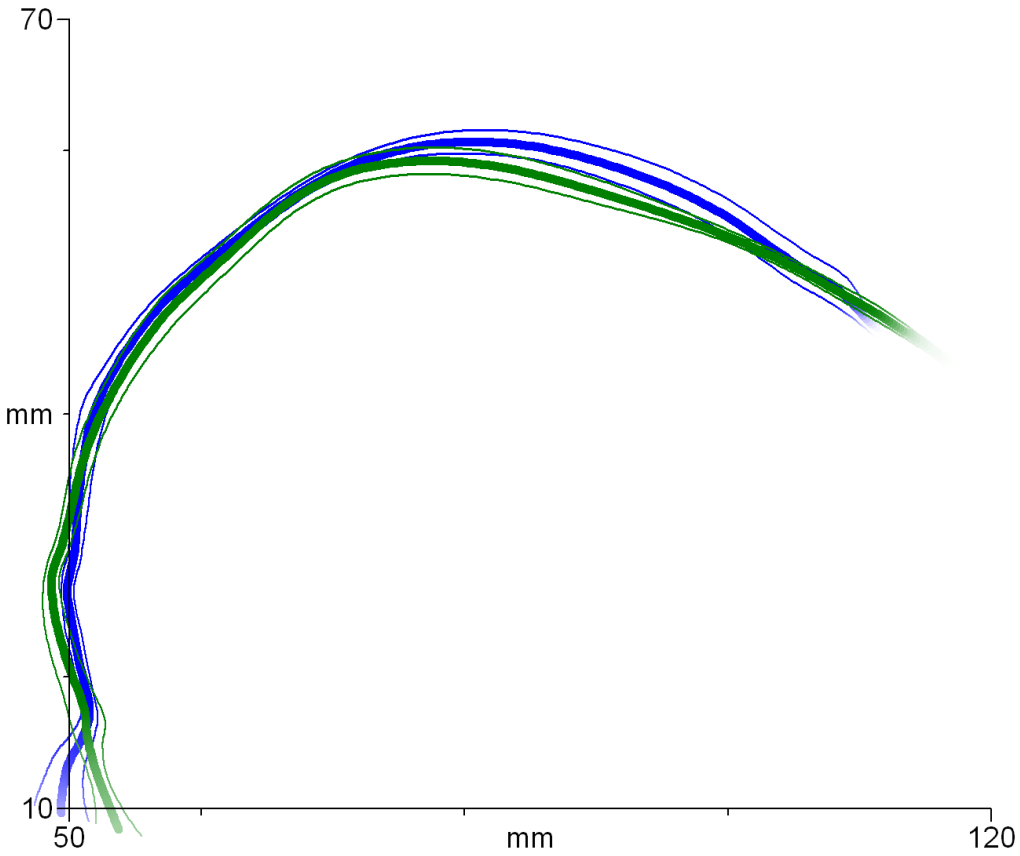


Figure 12 /a/

4.1.5 /ɔ/

More global differences are visible in this LOT & THOUGHT vowel, traditionally labelled as open-mid back. The Lombard version seems to show a global change, with retraction and lowering. There is a cross-over point somewhere in the upper pharyngeal / uvular area.

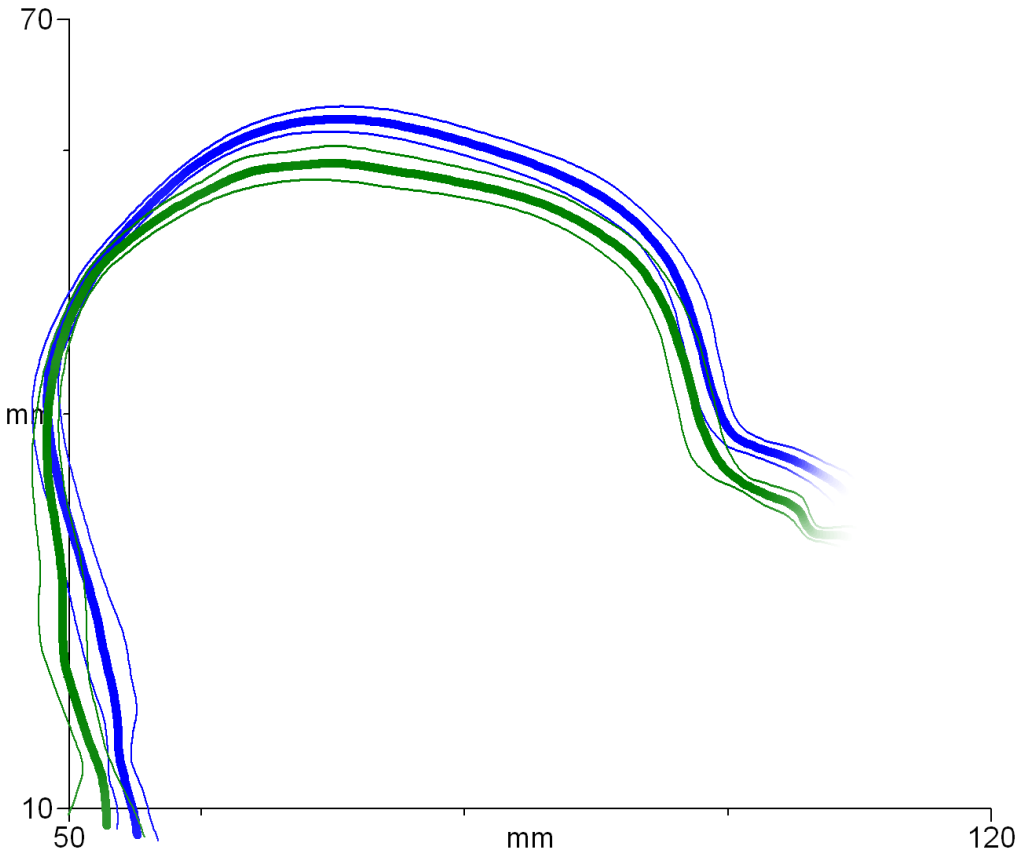


Figure 13 /ɔ/

4.1.6 /o/

Finally, the GOAT vowel also shows global changes of lowering and retraction. The cross-over point is a bit higher, perhaps uvular or velar, compared to /ɔ/, and the Lombard and neutral conditions appear to differ more. The floor of mouth in both these vowels is visible as the most anterior part of the "tongue" edge traced in the images – it is the lower part of the vocal tract section visible in the image so it seems appropriate to leave it in the images. The lowering of this feature may indicate a combination of tip location changes but more likely is due primarily to jaw lowering.

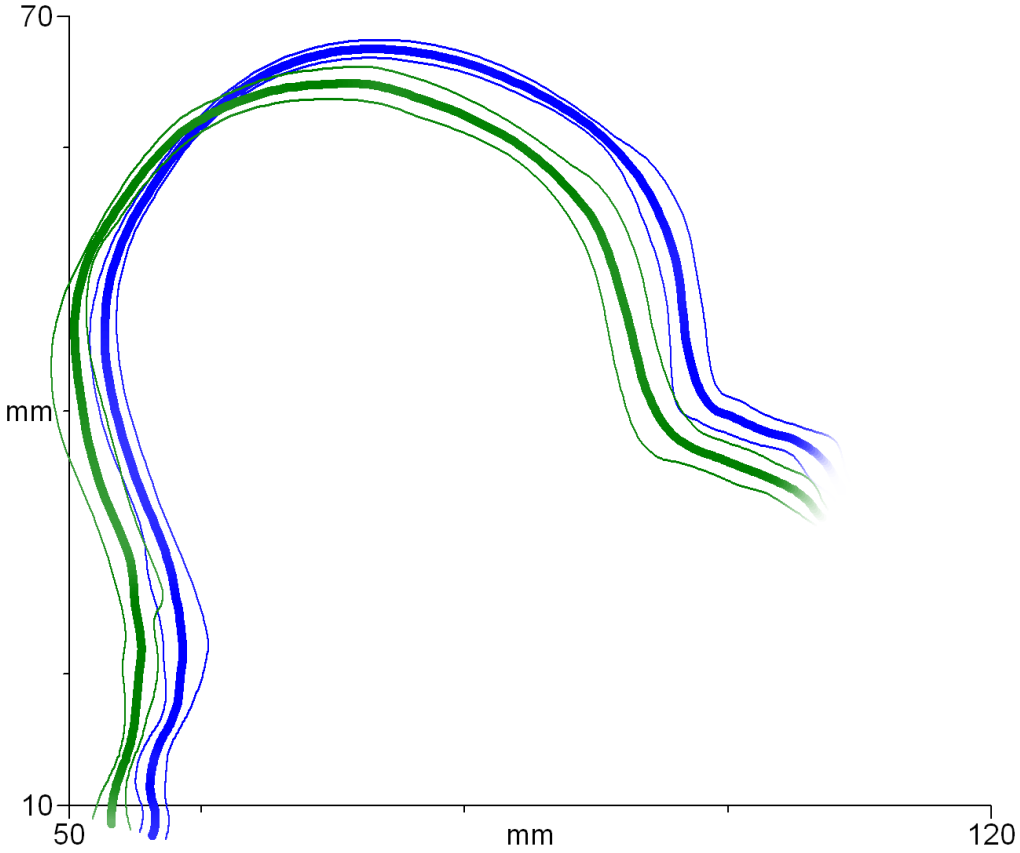


Figure 14 /o/

4.2 Quantitative differences in tongue surface shape and location

Some differences in the absolute position of the vowels relative to /i/ and the occlusal plane can be computed, along the lines of Scobbie et al (2013), but this is not the main interest here (see Appendix 4). Rather, this pilot study seeks to investigate the Lombard effect, and also to explore some methodological issues in quantification of differences between vowels, so we will focus on the Lombard effect below, within each vowel phoneme.

The difference between pairs of vowels involved multiple t-tests along fan-line radii, evaluation of confidence values for the mean tongue location in each location, evaluation of the number of adjacent significant t-tests, and the location and type of other non-significant t-tests. For example, as explained above, a cross-over by necessity gives a non-significant result in this type of analysis, even if a statistical test of surface tangent orientation/angle at that point would be significant. A significant difference (5 or more adjacent fans) was found for all vowels except /ʉ/, so therefore the quantificational analysis of tongue shape did not find a Lombard effect for this vowel, just the trend towards a change in shape similar to the significant changes in /i/ and /e/. The non-significant difference between the conditions can still be quantified, however, and is included in the descriptive statistics below, as a trend.

With non-significant /ʉ/ some assumptions are needed about how to quantify the size of this trend. There are two possible solutions. The first, conservative approach is based on the individually significant radial differences only: it quantifies just the distance between four adjacent significant anterior fan-line radii. This is what is presented below. The alternative is to include some relevant (non-adjacent or non-significant) fan-line values: specifically the single fan-line radius that is anterior to, and the three flanking non-significant values coincident with, the cross-over; and the single significant fan-line posterior to the cross-over (nine fan-lines in total). See footnote 1 for those results.

The results are given in Figure 15 in two measures. The first is the length of tongue surface (actually, tongue surface + floor of mouth surface if traced) judged to be involved in distinguishing the conditions (line series, in mm). The method used results in a larger value than just the length over which *significant* values were found, since cross-over and flanking non-significant (but confident) means are included. The non-significant trend for /ʉ/ is only for 1.5cm of surface length, which is indeed small.[1] The significant differences of at least 5 fan-lines (see Appendix 5) echo the Figures above. The vowel with the smallest significant Lombard effect is /e/, and the largest is /o/, which has nearly 9cm of surface over which a difference can be confidently posited.

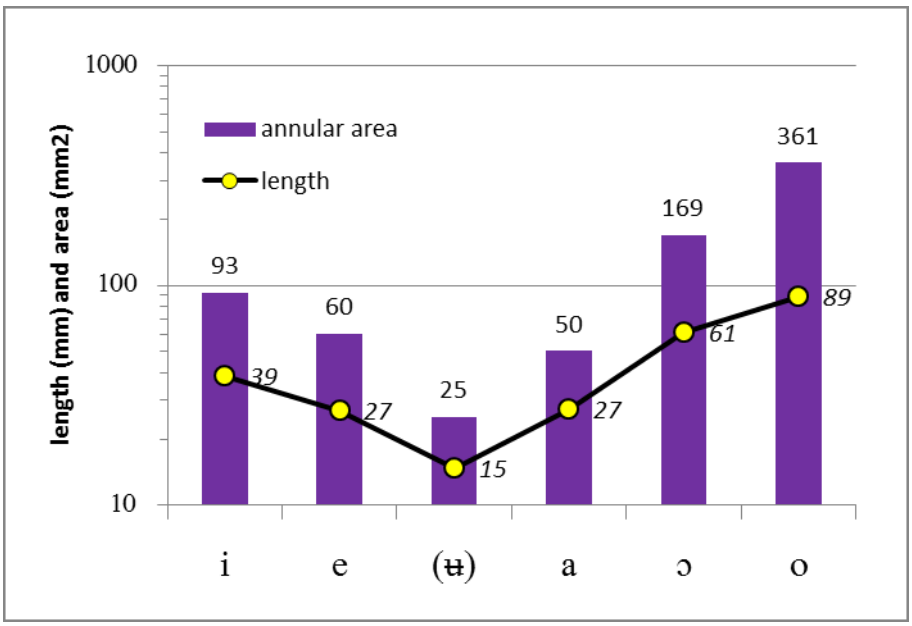


Figure 15 Articulatory Lombard effect size in mid-sagittal tongue sections, with a logarithmic y-axis to emphasise the comparable area-to-length results.

[1] The more radical solution, over 9 adjacent fan-lines, most obviously would result in a greater length (34mm) of tongue surface being considered (more than /e/ and /a/). The total area of difference would increase somewhat to 34mm^2. The maximum radial difference (1.2mm) would not change. Finally, the average area difference per mm of surface analysed would drop from 1.7mm^2 to 0.99mm^2, and the RMS difference would drop to 1.2mm.

The differences in area are roughly proportional to the square of the surface length to which they pertain. The inclusion of areas of cross-over, with by definition near-zero difference, mean however that the difference in area is a conservative measure when compared to the length over which the difference is found. The lack of difference at the cross-over or small flanking differences outside the larger, significant areas of difference, are included in the values reported here. The non-significant values for /ʉ/ show the potential sensitivity of the experimental design. More speakers would need to be analysed to work out if this vowel behaviour (whether is it is a lack of difference or a small difference) is systematic.

For the five other vowels, which have a larger absolute difference in area in Figure 15, when this extent of difference is normalised by dividing it by the length of difference, a more conservative value emerges (Table 1, Figure 16). This shows that each 1mm of tongue surface length in a region of difference has about 2mm of radial difference between the curves, on average, except in the case of /o/, with over 4mm. The vowel /ʉ/ which is not significant, has the lowest difference, on average, but it is not much less than the difference for /a/. The RMS difference calculated from the relevant fan-lines gives very similar results from linear measures, so is simpler to calculate.

Table 1 Difference between tongues: total area (mm^2) divided by overall length of the difference (mm), RMS difference (mm), and maximum radial distance (mm)

	i	e	(ʉ)	a	ɔ	o
Normalised area diff (mm^2/mm)	2.4	2.2	1.7	1.8	2.8	4.1
RMS radial diff (mm)	2.4	2.0	1.7	1.8	2.8	4.4
Maximum radial diff (mm)	3.5	3.4	2.1	2.2	3.7	9.8

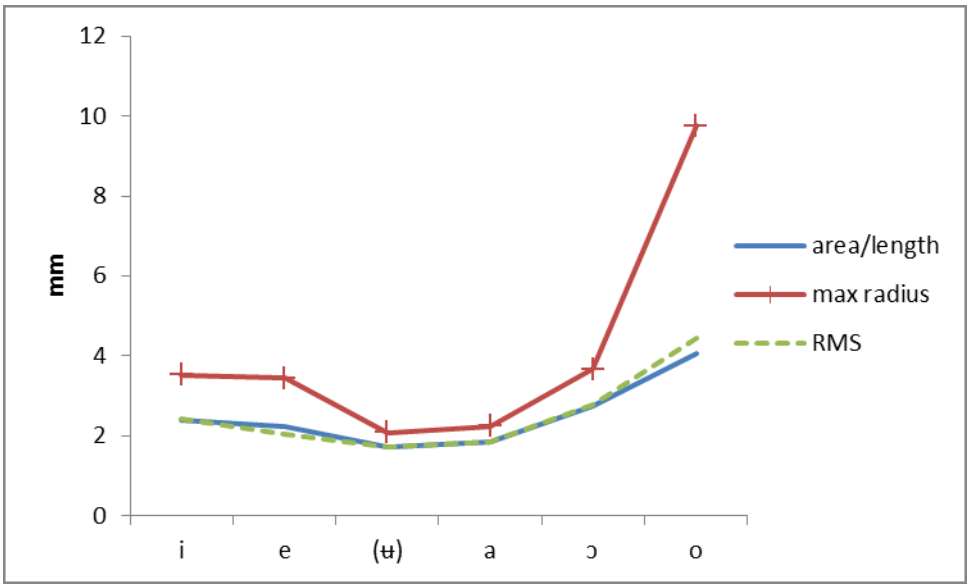


Figure 16 Two measures of the average (significant) difference between two curves compared to the single radial maximum difference

Since there is no rationale from speech production to choose a particular fan-line as a location in which difference is expected, we present no localised analysis. For the same reason, and because it is more susceptible to random effects, the single maximum radial difference does not appear to be a useful measure here.

4.3 Lip separation and compression

The vowels show a clear effect of increased lip opening for all vowels (Figure 17). Reminiscent of the non-significant trend in the tongue location, /ʉ/ has the smallest labial difference among the vowels. There is an apparent effect of vowel height, such that the lower, more open vowels do indeed have more widely separated lips, in both conditions.

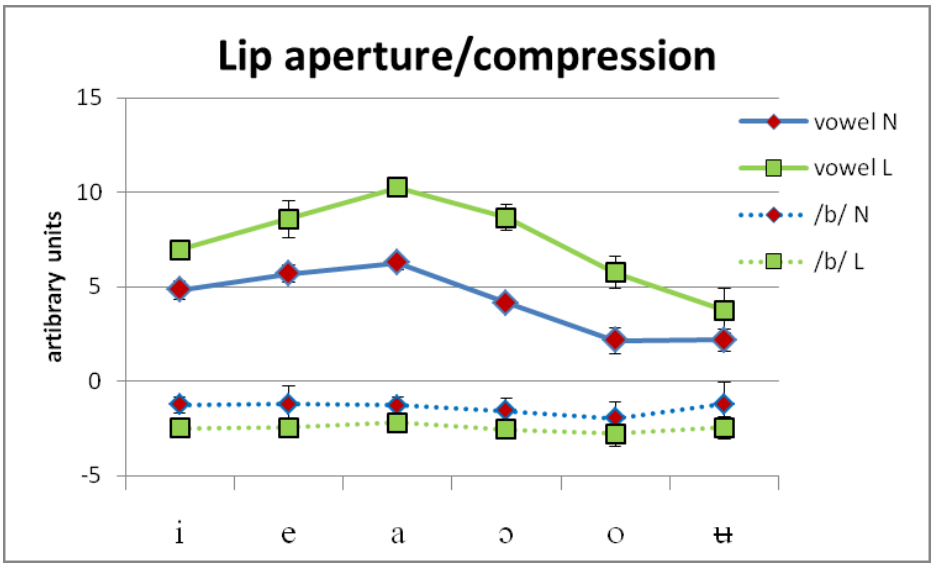


Figure 17 Lip separation, with zero value indicating approximately no separation and no compression. Dashed lines for /b/, solid lines for the vowels.

The effect of clear speech is that there is more active compression of the lips for /b/, and a wider opening for the vowel. Though not quantified due to the difficulty of segmenting /p/, this Lombard effect did not appear to be present for the /p/ in final position.

As well as the protrusion and opening effects, an attempt was made to quantify the degree of lip spreading and opening that was clearly visible in the images (Figure 18). The rounded vowels do show a slightly greater distance from the corner of the mouth to the lower lip's maximal protrusion in the Lombard condition than in the neutral condition, due it seems to extra rounding and opening (Figure 2). Spreading is however far more evidently exaggerated in the unrounded vowels.

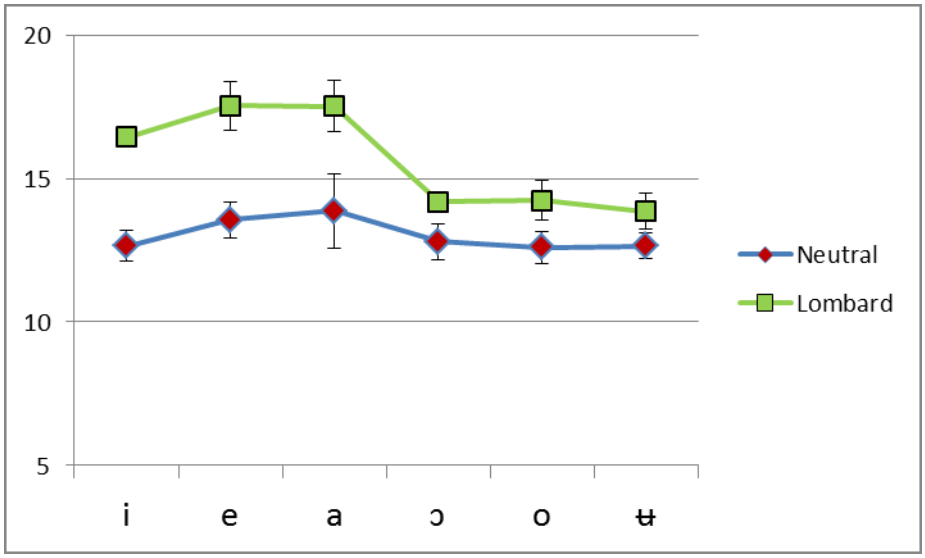


Figure 18 Lip spreading/opening as the mean distance from the lower lip's aperture tangent contact point to the corner of the mouth (aribitrary units).

4.4 Lip location

The common tangent line drawn to link the extremities of lips shows changes to both the overall distance (Figure 17) and the location (Figure 19 below). Qualitatively, the three unrounded vowels /i e a/ show some retraction in the Lombard condition rather than the increased protrusion evident in the rounded vowels /ɔ o ʉ/. All bar /ʉ/ seem to show some lowering of the lower lip and all appear to show raising of the upper lip, whether protruded or retracted.

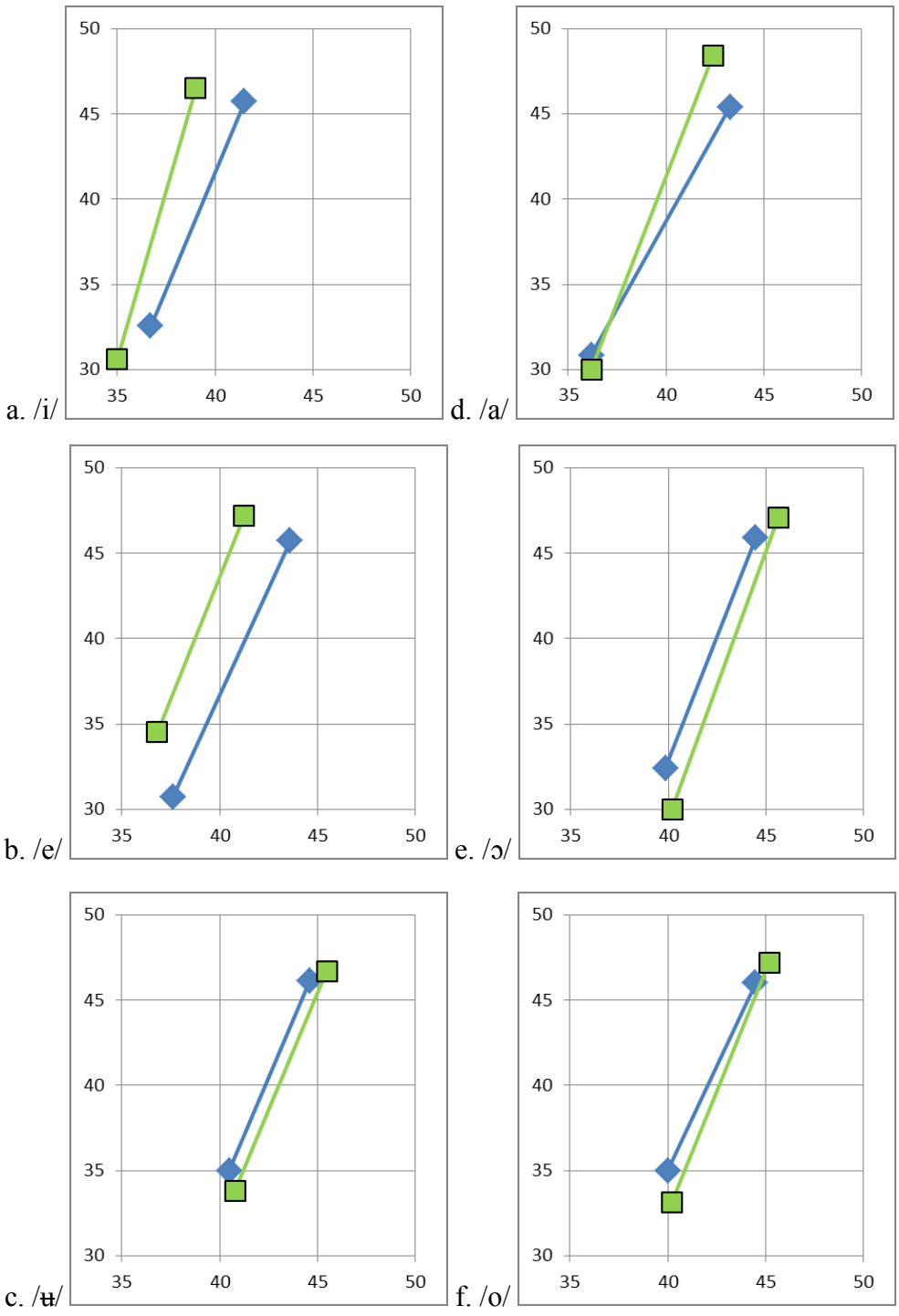


Figure 19 Location of lip tangent line (arbitrary units) in Neutral (blue diamond) and Lombard (green square) conditions. Anterior to right, from unrotated images.

4.5 Articulatory duration of labial closure

While it would be possible to measure the duration of the labial closure for initial /b/, it was not undertaken because in almost every production the speaker's lips were already closed pre-speech. Even though the lips before speech were still and uncompressed, so that the timepoint when compression or protrusion began could be seen could be quantified, it was hard to operationalise this manually, and so the initial /b/ was not measured. Measuring the articulatory duration of final /p/ was less problematic but on occasion the post-burst lip opening was very small or subtle. On the whole the /p/ closure was relatively easy to measure (Figure 20). However, the camera orientation and angle could have been better (see below): given the upward angle of the profile video, we think these measures are conservative, are likely to be too long.

We found no overall effect on /p/ duration. On average, the articulatory closure of final /p/ was 74ms (grand mean) in each condition. This is in contrast to the consistent (20% mean) reduction in acoustic silence corresponding to the final /p/, which reduced from 84ms in the Neutral to 67ms in the Lombard condition (Figure 5). The duration measures were equally variable, in that the average coefficient of variation of the 12 mean measures for articulatory measure of final /p/ was 20%, while the acoustic measure was 21%.

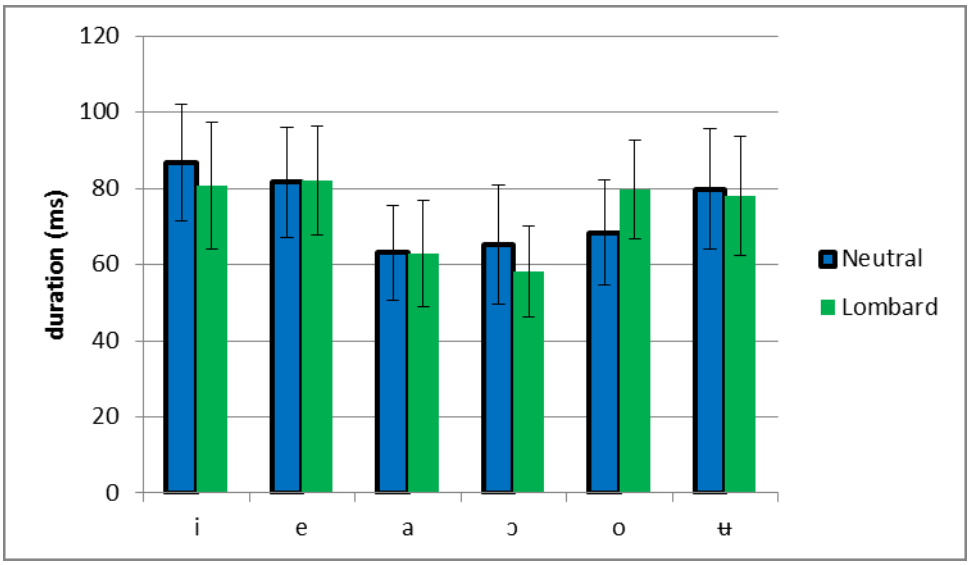


Figure 20 Articulatory duration of final /p/

5 Discussion and conclusions

Lombard speech is a type of clear speech elicited by the presence of environmental noise. Investigation of such speaker-controlled variation can provide insights into the nature of the phonology of a specific language as well as letting us probe general principles behind the complex relationships of: phonology to phonetics; articulation to acoutics, and both articulation and acoustics to perception. This pilot investigated the

prospects for using ultrasound tongue imaging and and lip video alongside more traditional acoustic analysis methods and asked how the tongue and lip targets were altered in an experimental Lombard speech condition in Scottish English. Noise that impaired a hearer's ability to identify a random vowel in a fixed structure CVC word (/b__p/) from a set of six, in a face-to-face communicative task, encouraged the speaker to enhance their CVC single word output. The mid-sagittal profile of their tongue and an approximately lateral view of the lips were captured in addition to acoustics, and analysed above.

The acoustic effects of the experiment seemed typical of Lombard speech, though the enhanced perceptibility of each word vis-à-vis the others in the materials used (or globally) cannot be tested due to clipping in the audio recordings. Nor can we explore the signal-to-noise ratio, the intensity of the speaker or level of the signal in audio or visual channels to the listener. Indeed, a host of relevant factors were not controlled. However, it was qualitively clear to observers and to the experimental listener at an impressionistic level that the Lombard condition speech in this binary condition task was clearer. Acoustically, intensity increased, F1 increased, duration increased and the vowel took up a greater proportion of the word. (Impressionistically, pitch excursion and pitch peaks also increased.) However, there was no change in F1 for the high front vowel /i/, and F2 increased slightly for /ɔ/ and paricularly /o/ (the de facto high back vowel of the system), which is why the vowel space did not increase in area in a general hyperspace effect (unlike Johnson et al., 1993) but just shifted in F1-F2 space.

In articulation, the Lombard versions of the vowels showed greater lip opening suggesting a lowered jaw and a more open vocal tract. However, the tongue did not appear lower for all vowels. The high front vowels /i/ and /e/ showed apparent tongue retration rather than lowering: a narrow contriction was by and large maintained in the palatal area, though it appears, surprisingly, to be more posterior than in the neutral condition, due to some tongue blade lowering. Surprisingly, /ʉ̟/ shows very little lingual change in the two conditions. Insofar as there is any subtle and statistically non-signficant changes, they appear similar to those seen for /i/ and /e/. There are two ways in which the probe location may have been affected, which would underestimate therefore the difference between the conditions. First, it may have been carried lower by jaw opening – but this would affect low vowels, which were indeed seen to have tongue lowering. More importantly, upward tensing of the tongue towards the palate could have been accompanied by expansion of the muscles downwards in the submental region under the chin, pushing the probe lower even without jaw lowering. This needs to be controlled for in future work.

We should note that /ʉ̟/ in this speaker's system is similar in some ways to /ʉ̟/ in other Scottish speakers studied previously (Scobbie et al., 2013). From the F1-F2 plot (Figure 7) we can see the vowel is central/front, which it typical. Another similarity is that /ʉ̟/ has a similar F1 to non-high vowels (most convincingly, /e/ in the neutral condition). If in the Lombard condition it behaves more as a high vowel in terms of F1, this might suggest that a clear-speech variation more closely reflects the historic phonological status of this vowel as /high/. However, it is stable in F2.

It would be interesting to explore more aspects of the linguistic situation here, and we cannot be sure whether the dialect mis-match betweens speaker and listener was important, let alone the effects of gender, task, context etc.

It would also be interesting to look at the visual information in the facial movements, not just reflecting the lip (and by extension jaw) articulation, but any indication within the mouth of tongue position. Ideally a front-facing (coronal plane) image would be collected, rather than the sagittal view of the lips here.

A final intriguing aspect of the dynamics of enhanced speech that should be explored in detail arises from the additional lip compression shown by /b/. This does not apply to final /p/, which is shortened and has a lax articulation. But neither /b/ nor /p/ varies in the materials, so there is no rationale for the enhancement of either consonant from the point of view of phonological clarification within the cohort of experimental materials. And none for the enhancement of just one of them. The explanation instead is likely to come from speech production planning and implementation.

One possibility that can explain the enhancement of /b/ is that even though a CVC word with fixed consonants appears symmetrical, if the vowel varies, then the whole word is holistically enhanced for clarity. Such whole word enhancement may be primarily concentrated in the initial CV portion. An alternative is that just the vowel segment is enhanced, but there is a spill-over or coarticulatory effect of enhancement into the initial C but not the final C. Nicolaidis (2012) reports that Schulman (1989) found that intervocalic labials were shorter in for loud speech. It would be well worth examining a range of different syllabic positions as well as considering the relationships between duration, the tenseness or degree of hyperarticulation in the consonant closure, and the speed of the movement in and out of the consonant. It would also be nice to compare the behaviour of an oral plosive (that requires intra-oral pressure) against a homorganic nasal stop (which does not).

One model that could be useful in understanding the syllabic patterns is Articulatory Phonology (Browman and Goldstein, 1986; Pouplier, 2011), since there is tighter coupling between an onset C and a following vowel (they are planned gesturally in-phase) than between a vowel and a following coda C (they are planned gesturally in an anti-phase, more sequential relationship). Whether it is the word as a whole or just the vowel that is the primary target of enhancement, then the asymmetry of gestural alignment in Articulatory Phonologywould provide a natural way to model such asymmetrical Lombard effects. Coproduction of enhanced V and neutral onset C could result in the sort of reactive labial compression seen here. Of course, this is contingent on better experiments that at least control the segmental materials, even in a simple expertimental protocol.

Clearly much more needs to be done to understand how the enhancement of some or all of the contrasts present in naturalistic contexts are actually achieved articulatorily, let alone how such changes are used to create certain acoustic (or visual) enhancements for the perceiver. We are certain, however, that techniques such as electropalography, ultrasound tongue imaging and video imaging are useful tools

for this essential task. Our goal will be to explore clear speech from the perspective of speech production, in relation to known aspects of both acoustics and perception.

Acknowledgements

We would like to thank the participants, and the other students who helped design and run this experiment as part of coursework. We express our special thanks to Steven Cowen and Alan Wrench for continuous technical support through the various stages of this study. Thanks also to Katerina Nicolaidis and other delegates at the Listening Talker conference (2012) for helpful discussion.

References

Alexanderson, S. and Beskow, J. (2013) Animated Lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. *Computer Speech and Language* **28**: 607-618.

Articulate Instruments Ltd (2008) *Ultrasound Stabilisation Headset Users Manual: Revision 1.4.*

Articulate Instruments Ltd. (2012) *Articulate Assistant Advanced User Guide: Version 2.14*.

Browman, C. and Goldstein, L. (1986) Towards an articulatory phonology. *Phonology Yearbook* **3**: 219-252.

Davidson, L. (2006) Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *JASA* **120**: 407-415.

Dreher, J.J. and O'Neill, J.J. (1957) Effects of ambient noise on speaker intelligibility for words and phrases. *JASA* **29**: 1320-1323.

Garnier, M., Bailly, L., Dohen, M., Welby, P. and Loevenbruck, H. (2006a) An acoustic and articulatory study of Lombard speech: global effects on the utterance. *Proceedings of Interspeech*, 17-22.

Garnier, M., Dohen, M., Loevenbruck, H., Welby, P. and Bailly, L. (2006b) The Lombard effect: a physiological reflex or a controlled intelligibility enhancement? *Proceedings of ISSP7*, 255-262.

Lallouache, M. T. (1991). *Un poste "Visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres.* Ph.D. ENSERG, Grenoble, France.

Lane H. and Tranel B. (1971) The Lombard sign and the role of hearing in speech. *JSHR* **14**: 677-709.

Johnson, K., Flemming, E., and Wright, R. (1993) The hyperspace effect: Phonetic targets are hyperarticulated. *Language* **69(3)**: 505-528.

Junqua, J.-C. (1996) The influence of acoustics on speech production: a noise induced stress phenomenon known as the Lombard reflex. *Speech Communication* **20**: 13-22.

Kim, J., Sironic, A., Davis, C. (2011) Hearing speech in noise: seeing a loud talker is better. *Perception* **40 (7)**: 853–862.

Lindblom, B (1990) Explaining Phonetic Variation: A Sketch of the H&H Theory. In W.J. Hardcastle and A. Marshall (eds.) *Speech Production and Speech Modelling*, 403-439. Springer.

Matthies, M., Perrier, P., Perkell, J.S. and Zandipour, M. (2001) Variation in anticipatory coarticulation with changes in clarity and rate. *JSLHR* **44**: 340-353.

Nicolaidis, K. (2012) Consonant production in Greek Lombard speech: an electropalatographic study. *Italian Journal of Linguistics* **24(1)**: 65-101.

Noiray, A., Cathiard, M.-A., Ménard, L., and Abry, C. (2011) Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in French and English speakers. *JASA* **129**: 340–349.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (1986) Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *JSHR* **29**: 434–446.

Pouplier, M. (2011) The atoms of phonological representation. In M. van Oostendorp, C.J. Ewen, E.V. Hume and K. Rice (Eds.) *The Blackwell Companion to Phonology*, 107-129.

Schulman, R. (1989) Articulatory dynamics of loud and normal speech. *JASA* **85**: 295-312.

Scobbie, J.M., Lawson, E., Cowen, S., Cleland, J, and Wrench, A.A. (2011) A common co-ordinate system for mid-sagittal articulatory measurement. *QMU CASL Research Centre Working Papers* **WP-20**.

Smiljanić, R. and Bradlow A.R. (2007) Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass* **3(1)**: 236–264.

Sumby, W. and Pollack,I. (1954) Visual contribution to speech intelligibility in noise. *JASA* **26**: 212–215.

Traunmüller, H. (1990) Analytical expressions for the tonotopic sensory scale *JASA* **88**: 97-100.

Uchanski, R. M. 2005. Clear speech. In D. B. Pisoni and R. Remez (Eds.) *The Handbook of Speech Perception*, 207–235.

Vatikiotis-Bateson, E., Barbosa, A.V., Chow, C.Y., Oberg, M., Tan, J., Yehia, H.C., (2007) Audiovisual Lombard speech: reconciling production and perception. *Proceedings of AVSP 2007: International Conference on Auditory Visual Speech Processing*. 41–45.

Wassink, A., Wright, R. and Franklin, A. (2007) Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. *Journal of Phonetics* **35**: 363-379.

Wrench, A.A. and Scobbie, J.M. (2006) Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. *Proceedings of the 7th ISSP*, 451-458.

.

Appendix 1 – Acoustic Duration

Mean	Vowel		Rime ("Word")		V/R Ratio	
	Neutral	Lombard	Neutral	Lombard	Neutral	Lombard
i	87	128	206	255	42%	50%
e	145	186	255	306	57%	61%
a	137	205	274	337	50%	61%
ɔ	121	173	230	305	53%	57%
o	135	181	234	290	58%	62%
ʉ	104	131	217	243	48%	54%
Grand mean	121	167	236	289	51%	57%

St. Dev	Vowel		Rime ("Word")		V/R Ratio	
	Neutral	Lombard	Neutral	Lombard	Neutral	Lombard
i	19	13	32	21	4%	4%
e	14	22	22	23	5%	4%
a	16	23	13	27	6%	4%
ɔ	15	28	24	24	5%	7%
o	11	29	14	36	4%	6%
ʉ	10	20	16	15	6%	6%

Mean	Silence "/p/ closure"		Transition		Combined	
	Neutral	Lombard	Neutral	Lombard	Neutral	Lombard
i	87	66	32	60	119	127
e	87	69	24	51	111	120
a	94	62	43	70	137	132
ɔ	79	67	30	65	109	132
o	74	68	25	41	100	109
ʉ	81	68	32	43	113	112
Grand mean	84	67	31	55	115	122

Appendix 2 – Acoustic Resonance

Hz	F1		F2	
Mean	Neutral	Lombard	Neutral	Lombard
i	337	325	2565	2582
e	417	504	2472	2395
a	885	967	1563	1541
ɔ	567	700	984	1035
o	416	516	910	1015
ʉ	398	443	1767	1761

Bark	F1		F2	
Mean	Neutral	Lombard	Neutral	Lombard
i	4.5	4.3	15.7	15.8
e	5.2	6.0	15.5	15.3
a	8.9	9.4	12.4	12.3
ɔ	6.5	7.6	9.5	9.8
o	5.2	6.1	9.0	9.7
ʉ	5.1	5.5	13.2	13.2

Appendix 3 – Coefficient of Variation (Bark)

Hz	F1		F2	
Mean	Neutral	Lombard	Neutral	Lombard
i	12.0%	3.4%	0.6%	0.3%
e	2.1%	4.5%	0.6%	1.1%
a	1.7%	1.1%	1.6%	0.7%
ɔ	4.1%	8.1%	3.4%	3.9%
o	4.2%	1.8%	2.3%	5.8%
ʉ	10.2%	6.2%	1.7%	2.0%